

Complete Genomics Data Overview and Workflows

January 28, 2011

Agenda

- Summary of projects
- Discussion of Data
 - Delivery
 - Storage
 - Processing
- Workflows
 - Germline
 - Tumor/Normal
- Future directions

Technology

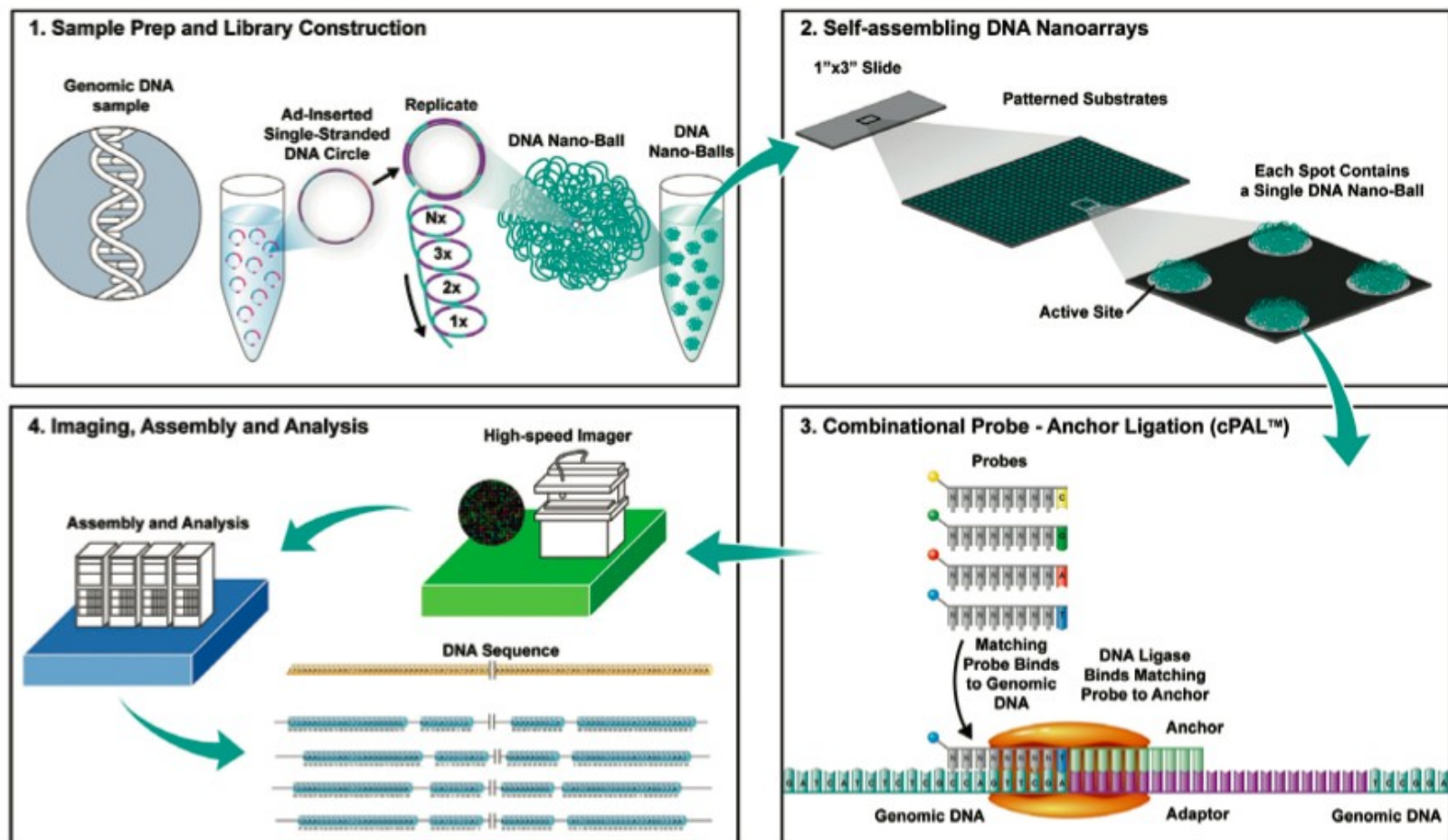


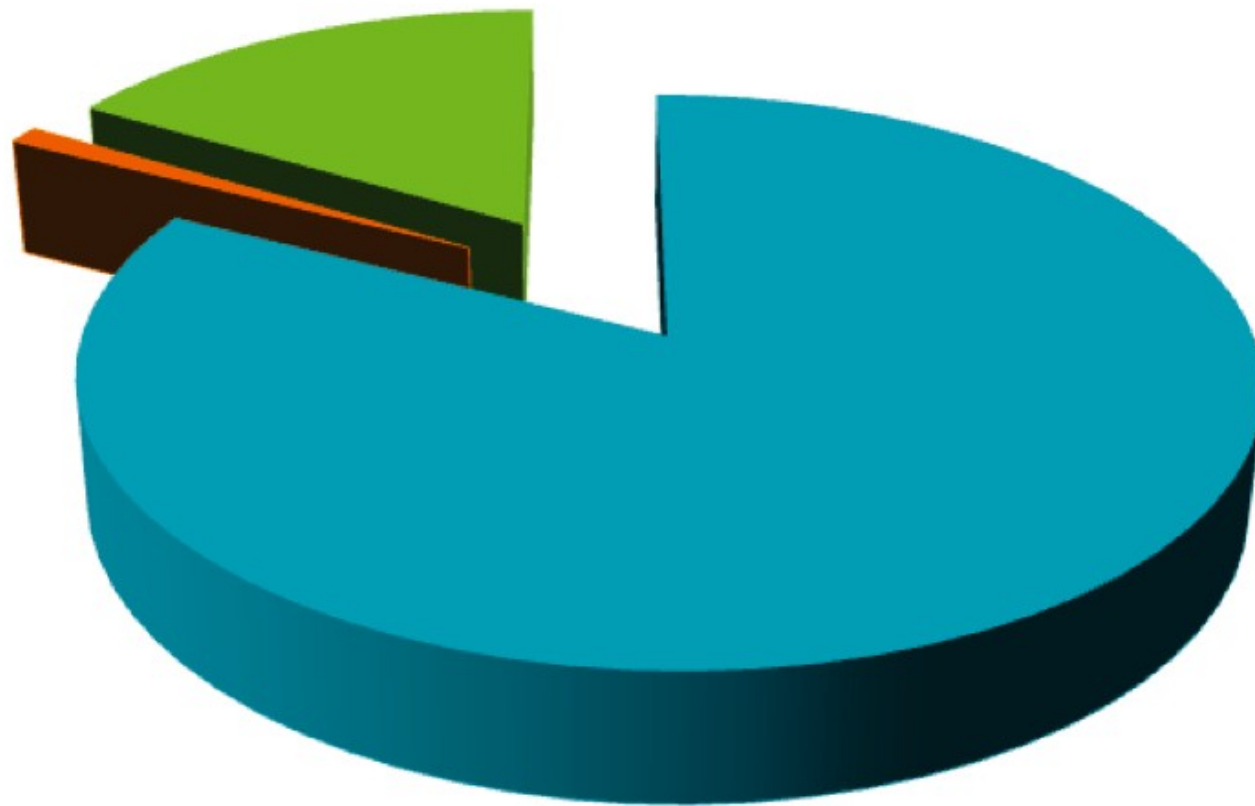
Figure 1 – Complete Genomics sequencing process

Data

- Delivery
 - Via USB/eSATA drive
 - Data kept by CGI for 30 days after delivery

Breakdown of Data Sizes

■ Reads & Mappings ■ Summary Reports ■ Assembly & Variations



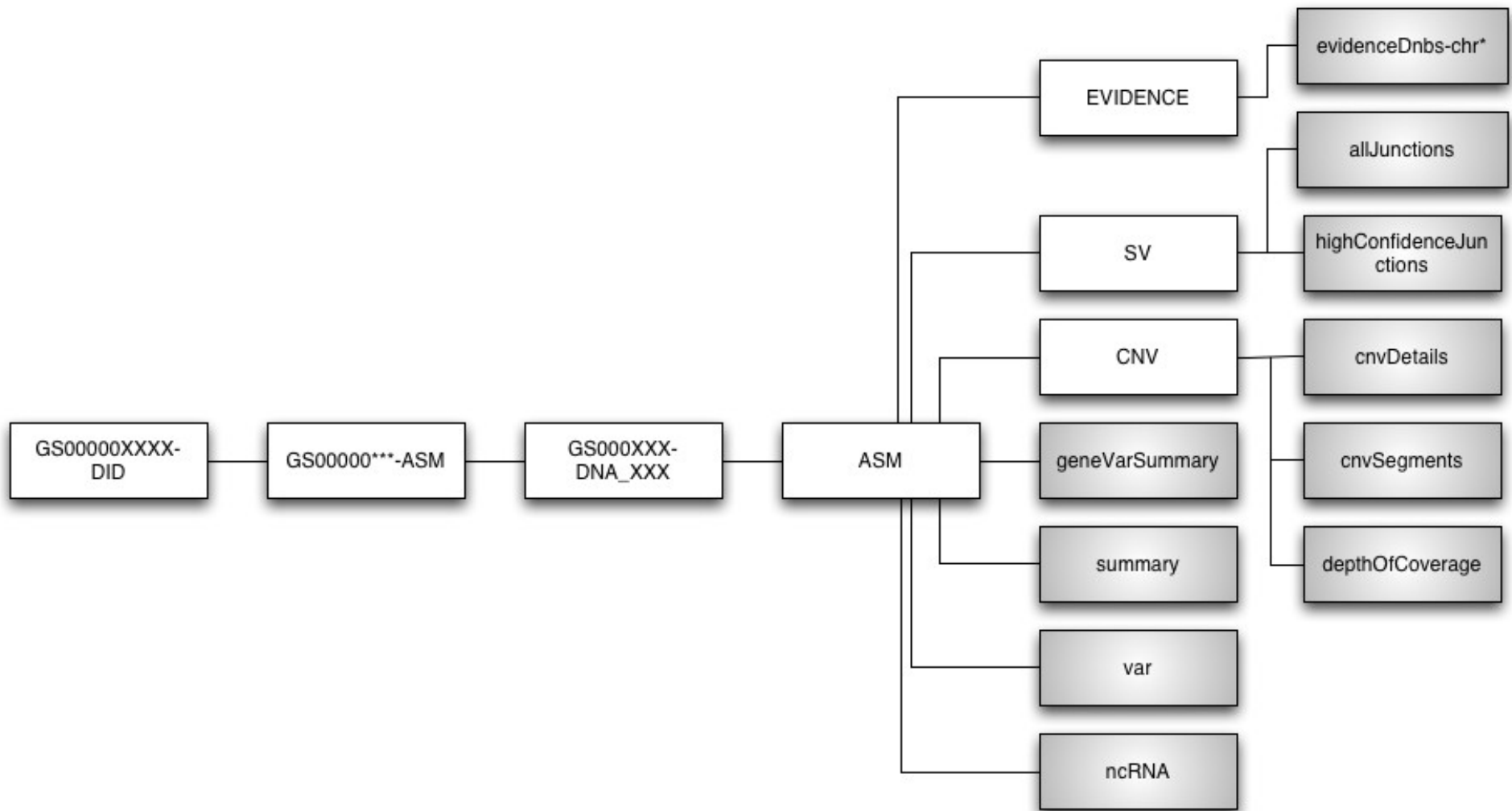
Data

- Delivery
- Storage
 - Sizes are LARGE
 - 400GB per sample as delivered to CCR
 - Should use 2-location backed-up storage
 - Not trivial to find such storage from NCI
 - Minimize:
 - Moving data around
 - Keeping multiple copies indefinitely

Data

- Delivery
- Storage
- Processing
 - Data are typically tab-delimited text files, so Excel can be useful for examining individual small files
 - Generally, command-line tools needed
 - MacOS and linux only supported operating systems, but Windows might work....
 - Some analyses (snppdiff) require large memory

Directory Structure



Data Summary

- Pause to look at data summaries....

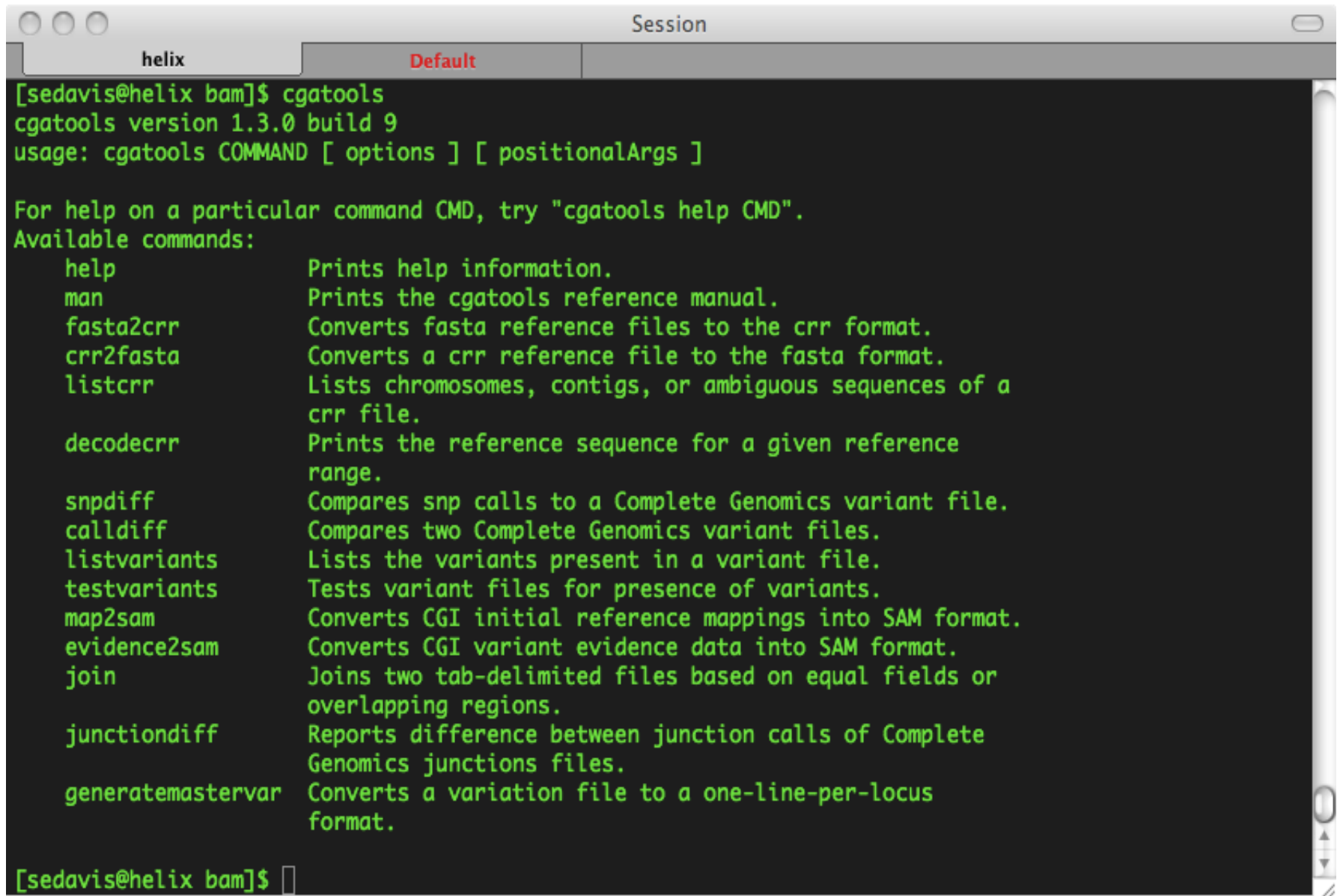
Workflows

- Tumor/Normal
 - *Copy Number*
 - *Structural Variation*
 - Annotated Somatic Variants
- Germline
 - List of annotated genotypes per individual, summarized into a single file that can be used for filtering

Tools

- Cgatools
 - Supported by Complete Genomics
 - Runs on linux and MacOS
 - Open Source
- Complete Genomics Toolkit (cgent)
 - Written in python
 - Does mainly file manipulation to enable visualization and interesting analyses
- Samtools, circos, annovar

Cgatools



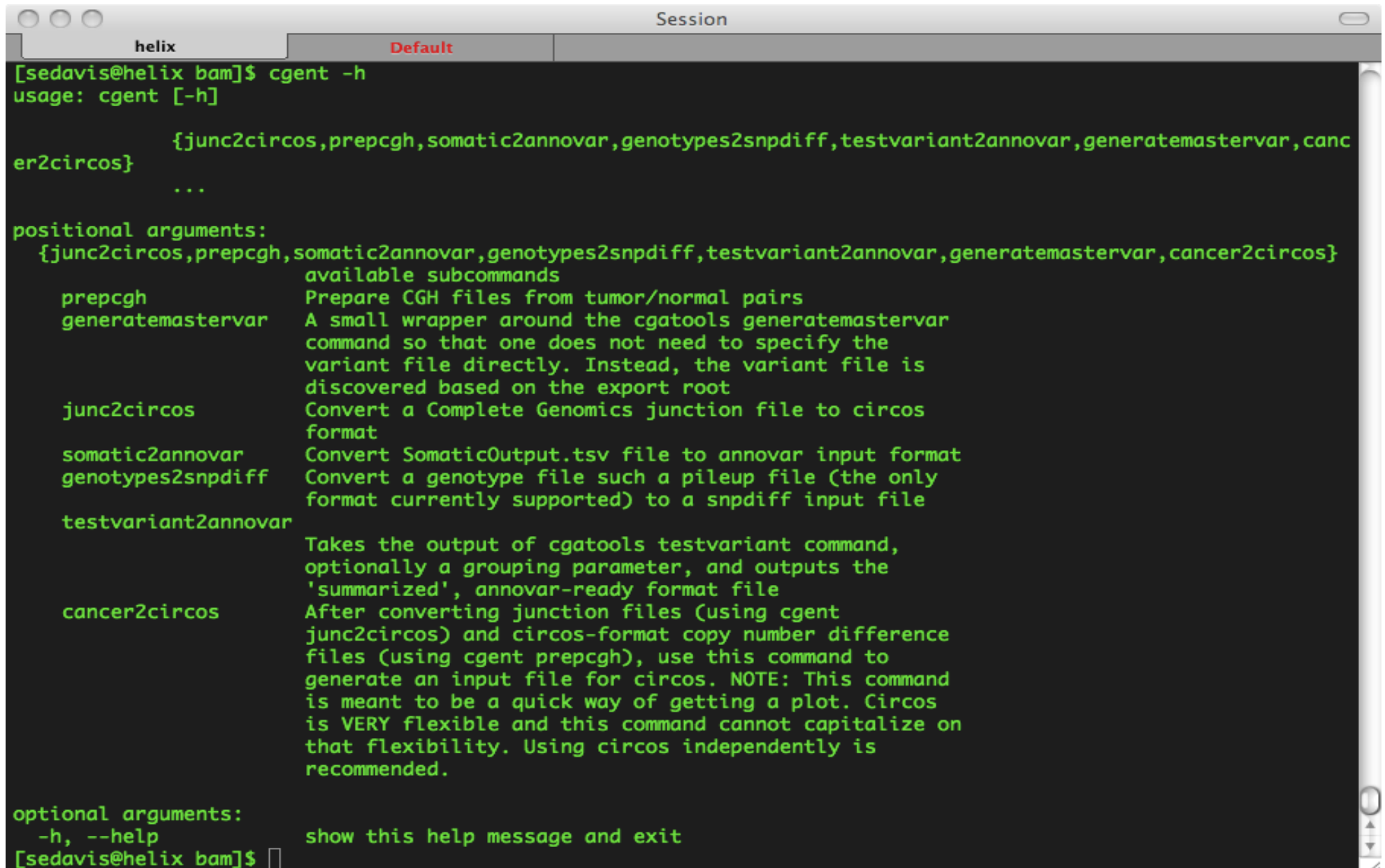
A terminal window titled "Session" with tabs for "helix" and "Default". The terminal shows the execution of the "cgatools" command, displaying its version (1.3.0 build 9) and usage instructions. It then lists available commands and their descriptions.

```
[sedavis@helix bam]$ cgatools
cgatools version 1.3.0 build 9
usage: cgatools COMMAND [ options ] [ positionalArgs ]

For help on a particular command CMD, try "cgatools help CMD".
Available commands:
  help          Prints help information.
  man           Prints the cgatools reference manual.
  fasta2crr     Converts fasta reference files to the crr format.
  crr2fasta     Converts a crr reference file to the fasta format.
  listcrr       Lists chromosomes, contigs, or ambiguous sequences of a
                crr file.
  decodecrr     Prints the reference sequence for a given reference
                range.
  snpdiff       Compares snp calls to a Complete Genomics variant file.
  calldiff      Compares two Complete Genomics variant files.
  listvariants  Lists the variants present in a variant file.
  testvariants  Tests variant files for presence of variants.
  map2sam       Converts CGI initial reference mappings into SAM format.
  evidence2sam  Converts CGI variant evidence data into SAM format.
  join          Joins two tab-delimited files based on equal fields or
                overlapping regions.
  junctiondiff  Reports difference between junction calls of Complete
                Genomics junctions files.
  generatemastervar Converts a variation file to a one-line-per-locus
                format.

[sedavis@helix bam]$
```

cgent

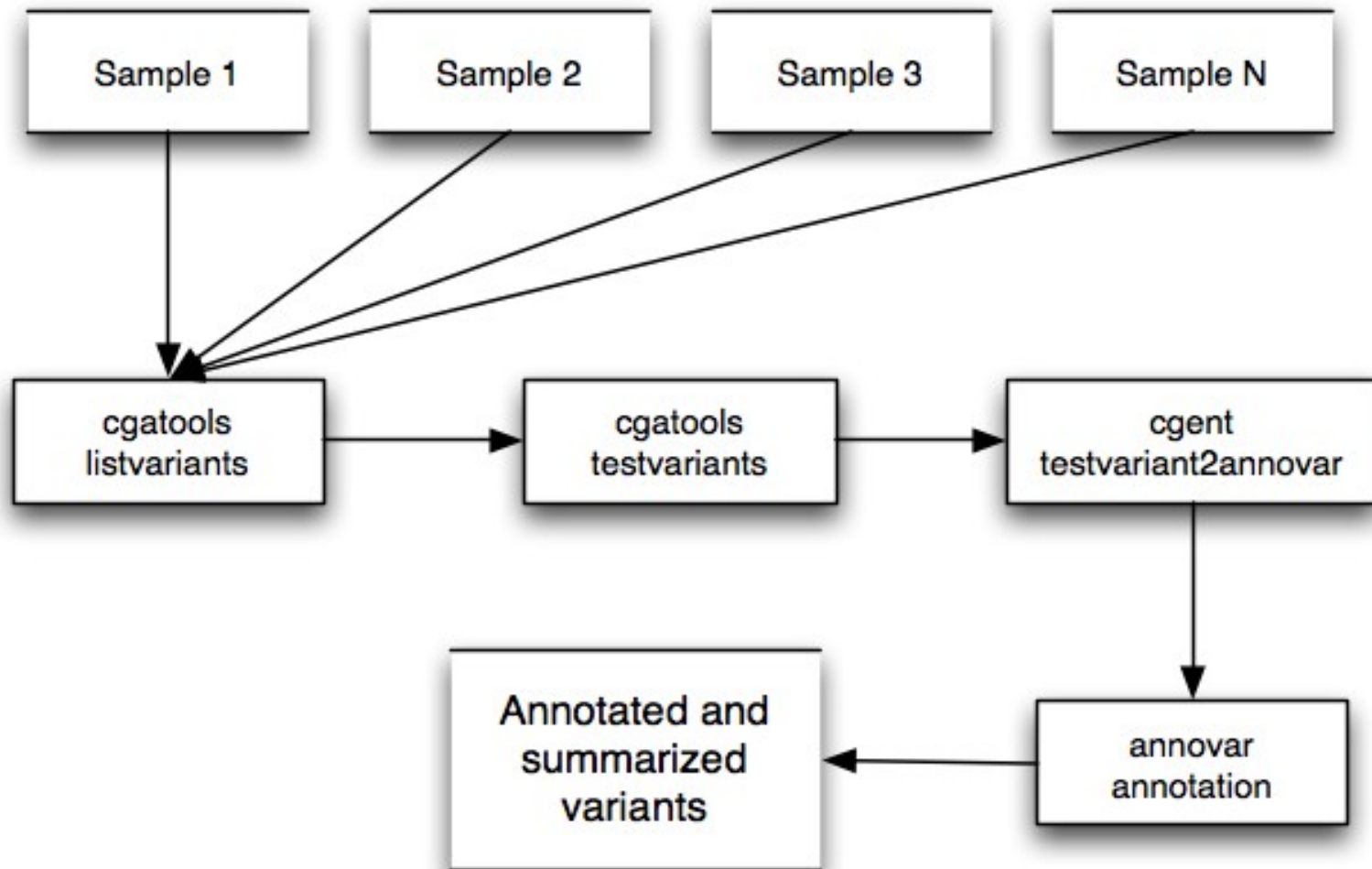


```
[sedavis@helix bam]$ cgent -h
usage: cgent [-h]
               {junc2circos,prepcgh,somatic2annovar,genotypes2snpdiff,testvariant2annovar,generatemastervar,cancer2circos}
               ...

positional arguments:
  {junc2circos,prepcgh,somatic2annovar,genotypes2snpdiff,testvariant2annovar,generatemastervar,cancer2circos}
    available subcommands
    prepcgh                Prepare CGH files from tumor/normal pairs
    generatemastervar      A small wrapper around the cgateools generatemastervar
                           command so that one does not need to specify the
                           variant file directly. Instead, the variant file is
                           discovered based on the export root
    junc2circos            Convert a Complete Genomics junction file to circos
                           format
    somatic2annovar        Convert SomaticOutput.tsv file to annovar input format
    genotypes2snpdiff      Convert a genotype file such a pileup file (the only
                           format currently supported) to a snpdiff input file
    testvariant2annovar    Takes the output of cgateools testvariant command,
                           optionally a grouping parameter, and outputs the
                           'summarized', annovar-ready format file
    cancer2circos          After converting junction files (using cgent
                           junc2circos) and circos-format copy number difference
                           files (using cgent prepcgh), use this command to
                           generate an input file for circos. NOTE: This command
                           is meant to be a quick way of getting a plot. Circos
                           is VERY flexible and this command cannot capitalize on
                           that flexibility. Using circos independently is
                           recommended.

optional arguments:
  -h, --help            show this help message and exit
[sedavis@helix bam]$
```

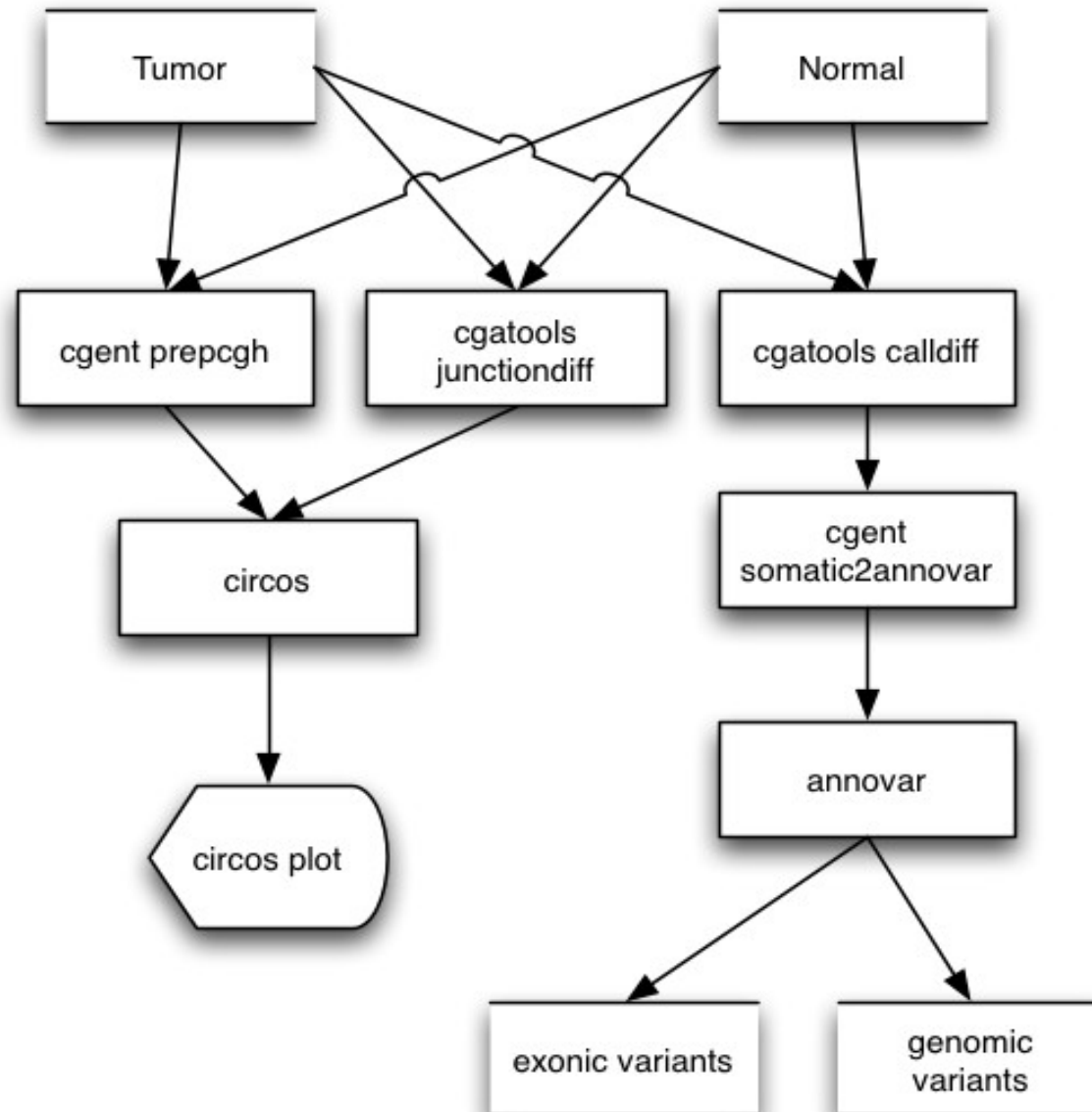
Germline Workflow

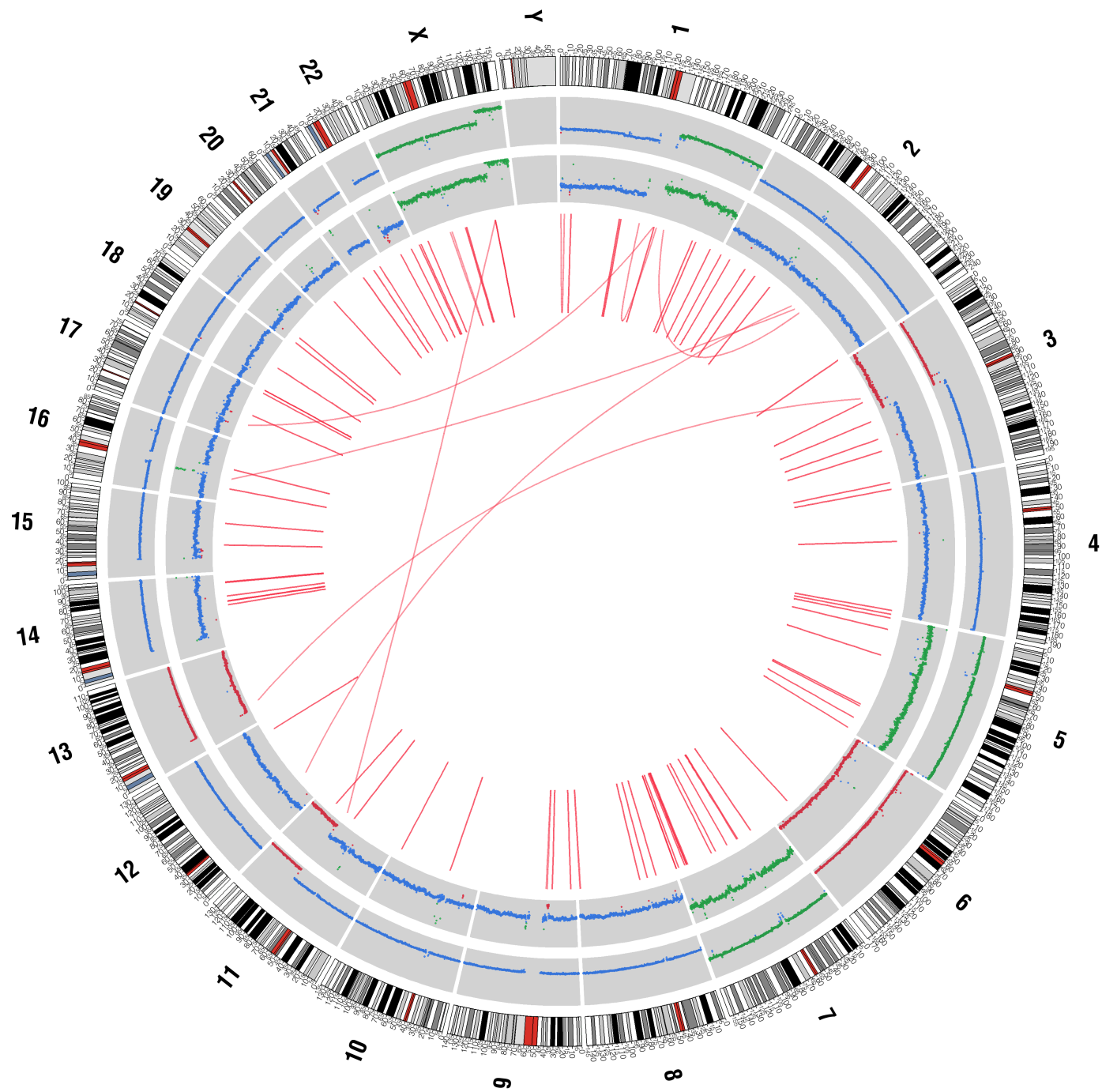


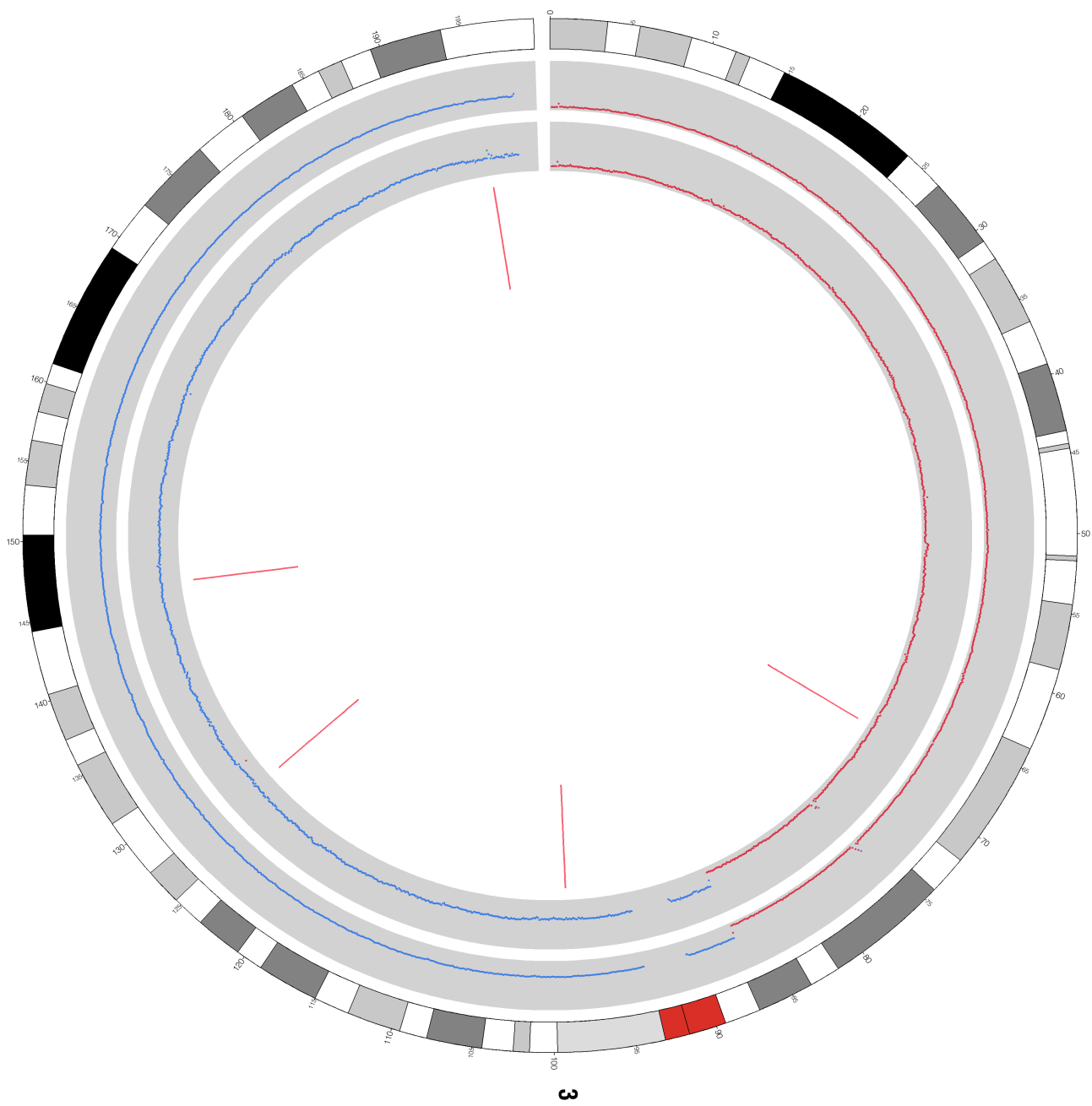
Germline Workflow

- Output
- Future directions
 - Be “smarter” about inheritance framework
 - Further refinements of comparison to other data types (exomes, snp arrays, RNA-seq)

Tumor/Normal Workflow



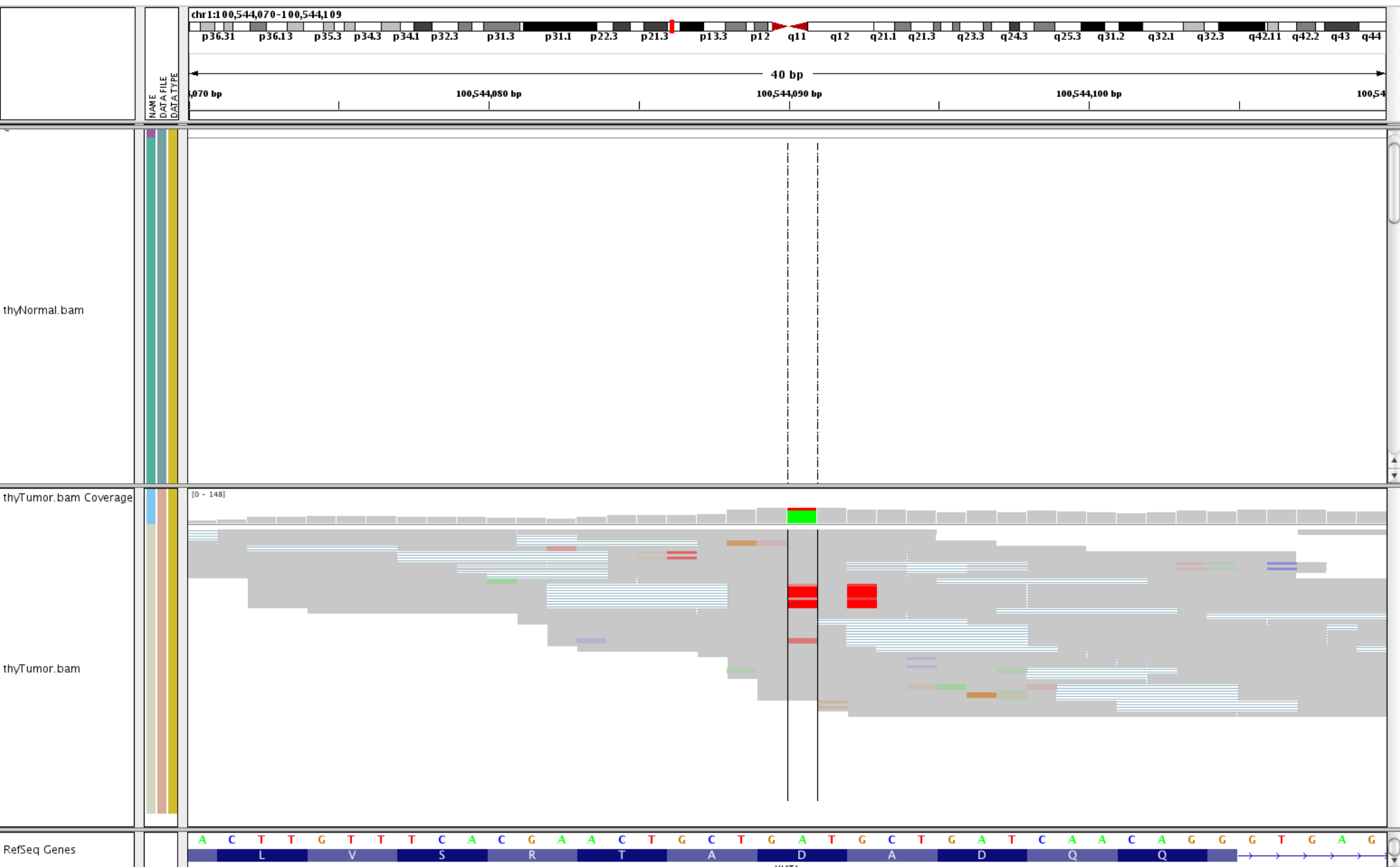


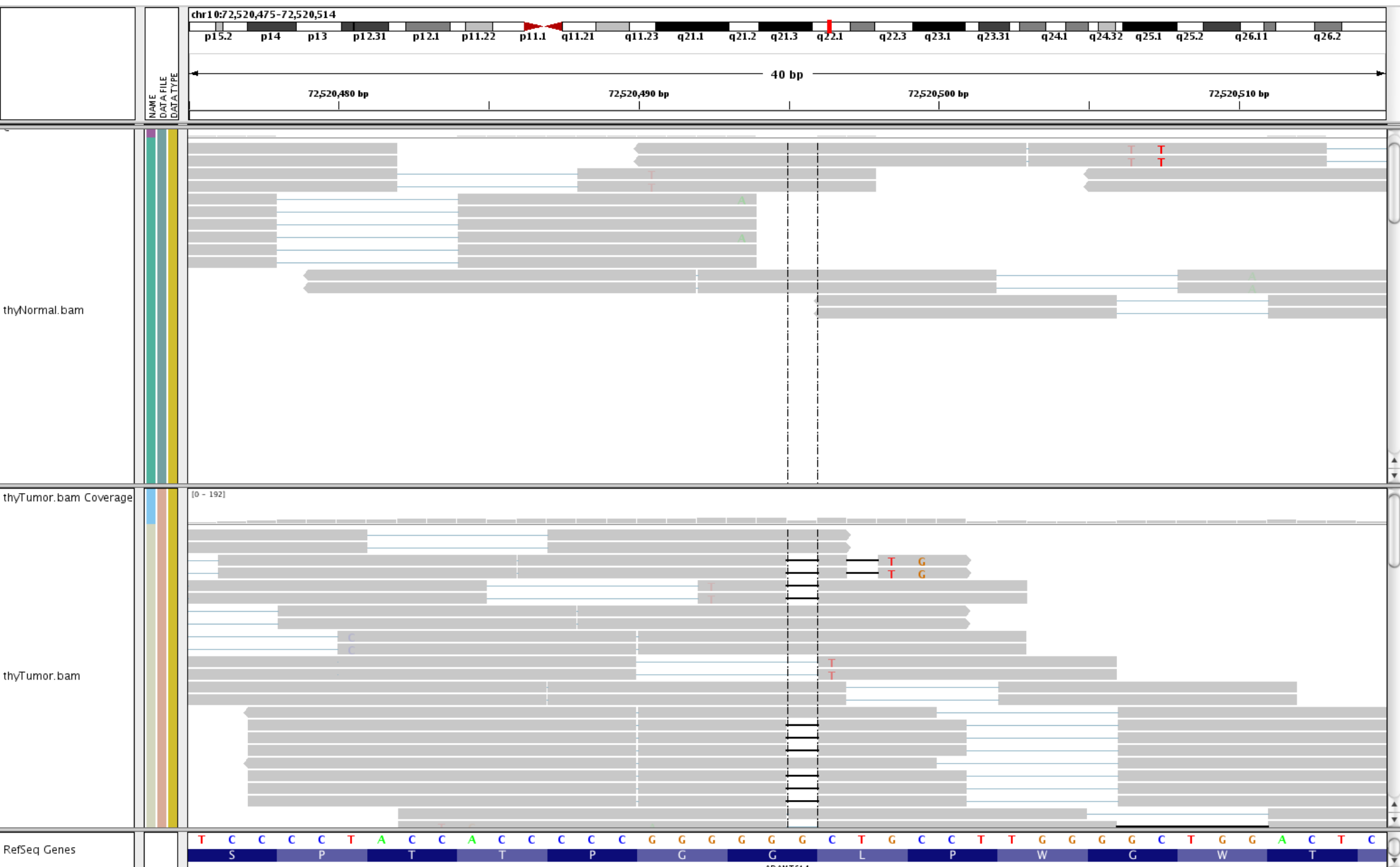


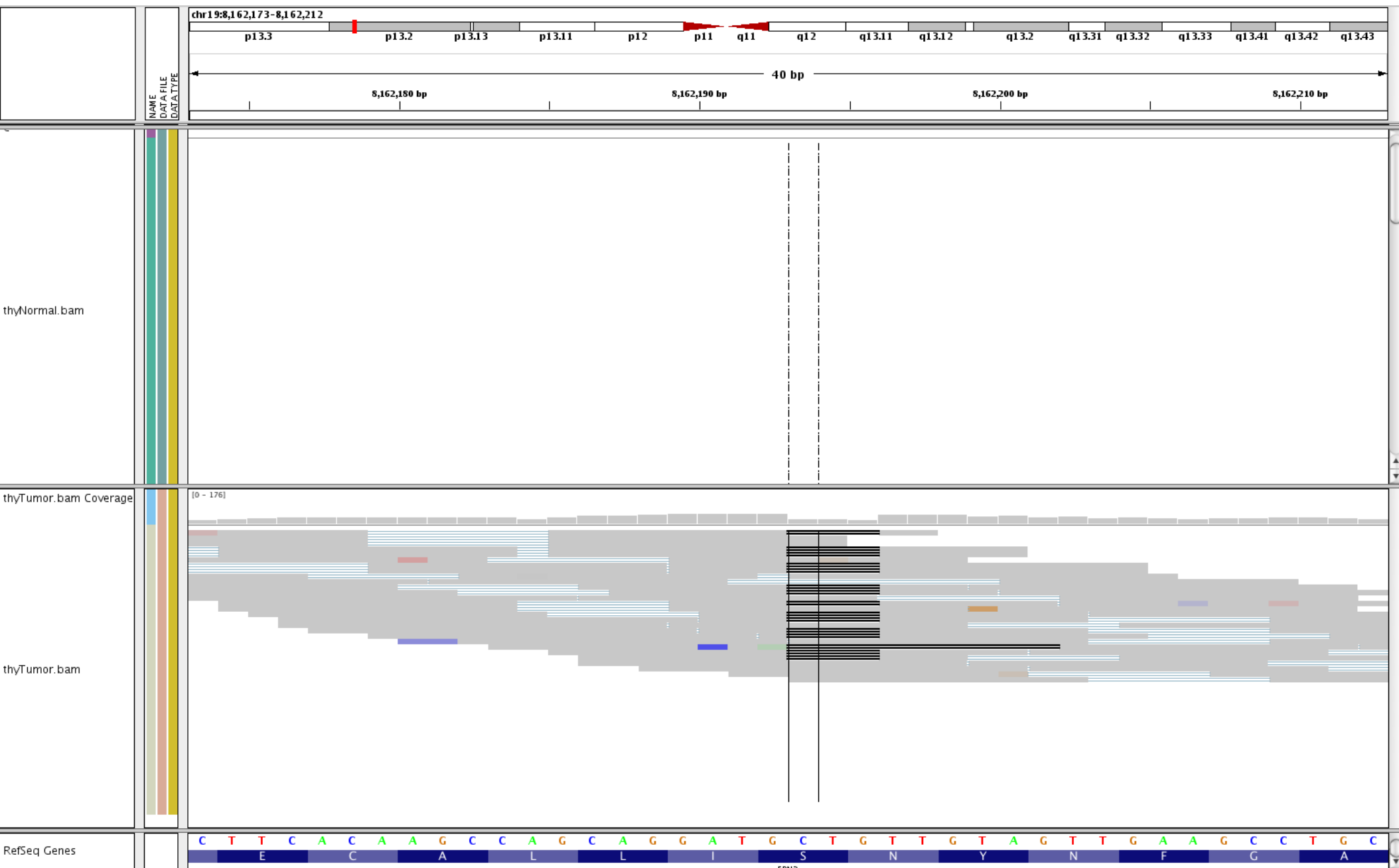
3

Tumor/Normal Workflow

- Text output example
- Future work
 - improve resolution of copy number (currently 100k)
 - Somatic variation list completion
 - Structural variation impact
 - Gene fusions
 - Overlap regulatory regions
 - Agreement with RNA-seq or mate-pair sequencing on Illumina







Open Discussion

- Storage
- Quality of data
- Pragmatic questions
 - Whole-genome sequencing?
 - Benefits of CGI versus Illumina?
-